Election Data

Scope of Project: find primary and general election results (from 2016 United States election) at the **county level** to map using GIS

Challenge: there is no single agency or organization that collects this information - we had to build a dataset from scratch

Questions:

- How do we collect data?
- What does it look like in its original form?
- What do we do to produce clean data?

Election data continued...

Data Collection:

- Searched state Board of Elections websites for data sets.
- Data is presented in a variety of formats: Excel, CSV, PDF, text files. Some states did not have a downloadable file available.
- The original data format dictated the amount of work necessary to clean the data. Our end goal was a CSV file for each state, with standardized name fields for candidates, number of registered voters, and voter turnout.

Election data - our end goal:

County	Clinton_Hillary	Stein_Jill	Johnson_Gary	Trump_Donald
ALLEN	1433	121	229	3651
ANDERSON	672	81	151	2435
ATCHISON	1989	163	316	4049
BARBER	286	28	78	1850
BARTON	1839	157	347	7888
BOURBON	1336	108	205	4424
BROWN	863	97	158	2906
BUTLER	6573	395	1415	19073
CHASE	316	31	45	969
CHAUTAUQUA	197	23	25	1236
CHEROKEE	2005	130	296	6182
CHEYENNE	181	20	29	1173
CLARK	120	20	36	825
CLAY	677	85	150	2891
CLOUD	761	74	161	2919
COFFEY	727	70	189	3050

Election data examples:

- Data in its (best) form: Ohio as an example of a file that required little work to be made into the form we needed.
- Data in an ugly form: Pennsylvania had no downloadable file, and the presentation of data on the website made it challenging to work with.
- Data in another ugly form: Tennessee as an example of some states / localities that only provide data as PDFs. This required us to convert files to Excel, then clean the files.

Coal Mine Accident Data

Scope of project: Analyze ~16,000 individual mining accident records to determine patterns by mine, mine operator, nationality of miners, occupation, etc.

The Data Files:

- Found as PDFs - one per letter of the alphabet
- Converted from PDF to Excel using Acrobat to have files we could clean (backwards conversion is FAR from perfect)
- Some files required more work than others (see cell A2)
- Our final spreadsheet, with clean data so that we could sort by various fields to look for patterns in the data

Census Data From Foreign Countries

Scope of Project: find subnational census data that shows population by race, ideally data is both current and historical → for selected countries in Latin America, in order to create map layers in GIS

Challenges: (1) There is no single agency or organization that collects this information - we had to build a dataset from scratch. (2) I don't speak or read Spanish or Portuguese - the language that much of the documentation was in. (3) Each country has its own census cycle, parameters for collecting data, etc. (lack of standardization) In some countries, data on race is not collected.

Census Data From Foreign Countries con’t

Data Collection:

- Searched country websites for census bureau or its equivalent. In some cases, referred to Wikipedia to find out when the last census was conducted.
- I had to rely on Google Translate to navigate sites that didn’t offer an English translation. I did not translate the words used to categorize race and/or ethnicity, because these carry cultural context.
- Data is presented in a variety of formats: Excel, CSV, PDF, text files.
- The original data format dictated the amount of work necessary to clean the data. Our end goal was a CSV file for each country, with population figures at a subnational level by race and/or ethnicity.

Lessons...

- Data is **messy**. Unless you're starting from scratch and building your own dataset, any data that you find online will likely require some clean up before it can be used or further analyzed.
- Clean-up can be a lengthy process, depending on the original form of the data and the size of your dataset.
- Data cleaning isn't necessarily difficult, especially after you learn some Excel formulas and other tricks that can help automate the process. But it does require patience and attention to detail.